DISCUSSION PAPER SERIES

No. 9880

THE CHOICE OF HONESTY: AN EXPERIMENT REGARDING HETEROGENEOUS RESPONSES TO SITUATIONAL SOCIAL NORMS

Rajna Gibson, Carmen Tanner
and Alexander F Wagner

FINANCIAL ECONOMICS and LABOUR ECONOMICS

Centre for Economic Policy Research

www.cepr.org

Available online at: www.cepr.org/pubs/dps/DP9880.php

ISSN 0265-8003

THE CHOICE OF HONESTY: AN EXPERIMENT REGARDING HETEROGENEOUS RESPONSES TO SITUATIONAL SOCIAL NORMS

Rajna Gibson, Swiss Finance Institute and University of Geneva
Carmen Tanner, University of Zurich
Alexander F Wagner, University of Zurich, Swiss Finance Institute, Harvard University, ECGI, and CEPR

Discussion Paper No. 9880
March 2014
Revised January 2015

Centre for Economic Policy Research
77 Bastwick Street, London EC1V 3PZ, UK
Tel: (44 20) 7183 8801, Fax: (44 20) 7183 8820
Email: cepr@cepr.org, Website: www.cepr.org

CEPR Discussion Paper No. 9880

March 2014; revised January 2015

ABSTRACT

The Choice of Honesty: An Experiment Regarding Heterogeneous Responses to Situational Social Norms*

We conduct a laboratory experiment in which we expose participants to situational social norms of approval or disapproval of lying. While participants on average conform to the situational pressure, the results highlight important differences in individual reactions. Situational norms crowd out intrinsic preferences for truthfulness; conversely, these preferences support resistance against "bad" norms. The extent and direction of the interaction of individual characteristics with situational norms and with economic incentives shed light on why people act truthfully. Out of several possible explanations, self-signaling under situational pressure provides the most convincing account of the evidence from the experiment.

JEL Classification: C91, G02, G30 and M14
Keywords: Crowding-out, honesty, norm conformity, protected values, self-signaling, situational social norms

Rajna Gibson
University of Geneva
30 Bd Post d-Arve
CH-1211 Geneva 4
SWITZERLAND

Email: rajna.gibson@unige.ch

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=111969

Carmen Tanner
Department of Banking and Finance
University of Zurich
Plattenstrasse 32
CH-8032 Zurich
SWITZERLAND

Email: carmen.tanner@bf.uzh.ch

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=178125

Alexander F Wagner
Department of Banking and Finance
University of Zurich
Plattenstrasse 14
CH-8032 Zurich
SWITZERLAND

Email: alexander.wagner@bf.uzh.ch

For further Discussion Papers by this author see:
www.cepr.org/pubs/new-dps/dplist.asp?authorid=166194

*We thank the Research Priority Program Finance and Financial Markets of the University of Zurich, the NCCR FINRISK, the Swiss Finance Institute, the Swiss National Science Foundation (Grant PP001-102845), and the ERC (FP7/2007-2013, grant agreement 249415 RMAC) for financial support. We thank Nicolas Berkowitsch and Elisabeth Proehl who both provided excellent research assistance, and Marcel Morf, who programmed the first version of the code for the experiments. We thank seminar participants at UC San Diego, the University of Lugano, and the University of Zurich for comments. We are grateful to Björn Bartling, Michael Brennan, Uri Gneezy, Michel Maréchal, Abraham Ravid, Jean-Charles Rochet, Joel Sobel, Janna Ter Meer, and Roberto Weber for comments, and to Peter Bossaerts, Chiddi Chidambaran, Alain Cohn, François Degeorge, Ernst Fehr, Gerlinde Fellner, Susanne Neckermann, Mei Wang, and Rainer Winkelmann for discussions on the general topic of this paper.

Submitted 16 February 2014; revised 22 July 2014 and 27 January 2015

1 Introduction

While it has long been known that situational norms and social pressure guide human action in direct and meaningful ways (Sherif, 1936; Asch, 1956; Milgram, 1974; Cialdini, Kallgren, and Reno, 1991), it is only recently that policymakers and managers are seeking to actively harness this power.[1] For example, in environmental policy, governments are increasingly turning to norm-based messages of inducing environmentally friendly behavior (e.g., Ferraro and Price, 2013). Tax authorities are also considering to enhance tax compliance by influencing social norms (see Luttmer and Singhal (2014) for a survey). Corporations, too, use social-norm based policies (such as codes of conduct) to foster ethical behavior (see Kaptein and Schwartz (2007) for a survey), aware of the fact that corporate culture and identity can greatly influence such behavior (e.g., Cohn, Fehr, and Maréchal, 2014) and that situational pressure can induce contagion effects both within and among groups (e.g., Gino, Ayal, and Ariely, 2009; Innes and Mitra, 2013). Martin (2012) argues that "every manager's tool kit should include an understanding of the power and ethical uses of social norms."

But little systematic evidence is available on individuals' heterogeneous responses to situational norms. This is surprising, given that the results of several theoretical papers depend on the conjecture that individuals respond differently to social norms. See, for example, Sliwka (2007), Fischer and Huddart (2008), and Huck, Kübler, and Weibull (2012). Ostrom (2000) and Andreoni and Bernheim (2009) also note that individuals care heterogeneously about norm conformity. Kimbrough and Vostroknutov (2014) indeed document substantial differences among individuals in their propensities to adhere to social norms, and argue that findings in games involving pro-social behavior can be

[1] Situational norms are temporarily made salient and are more transitory than long-standing internalized norms that are uniformly in force at all times and in all situations. They do not need to be internalized to be effective, but work as a result of social or group pressure; they can be relatively easily manipulated by managers and policy makers. We use the terms *situational social norms* and *situational pressure* interchangeably. Other terms could be used to describe this force: *normative influence* (Deutsch and Gerhard, 1955); *exhortations*; or *social norms*, which generally refer to injunctions on behavior that are sustained by the threat of social disapproval or penalties (Elster, 1989).

explained by these tendencies.

This paper provides further evidence of this heterogeneity. Our main contribution is to illuminate its sources. Rather than considering situational pressure and individual preferences separately, our analysis considers their interaction.

We study individuals' choices to act honestly. Understanding the sources of honesty is important because it may conversely also shed light on the drivers of personal misconduct (such as tax evasion and fraud) and corporate misconduct. Moreover, there is a clear theoretical framework for what kind of heterogeneous responses to situational norms one might expect. In particular, much like extrinsic incentives may *crowd out* intrinsic motivations, for example, to work hard, or to help others (see Bowles and Polanía-Reyes (2012) and Gneezy, Meier, and Rey-Biel (2011) for surveys), it is plausible that social norms crowd out intrinsic preferences for truthfulness. A large array of research has argued that some individuals experience intrinsic costs of lying, and some proxies for such lying costs exist that can be used in empirical work.[2] However, the interaction of these intrinsic costs of lying with situational norms has not yet been experimentally studied.

We conduct a laboratory experiment in a concrete context: accounting earnings management. In the experiment, earnings management is designed to be a form of lying, which is defined as making "a statement that one knows to be false" (Grover, 2005).[3] Specifically, participants are cast in the role of CEOs and are told the truthful level of earnings. However, they are informed that they can legally announce higher earnings and receive higher bonuses. The experiment is set as an anonymous one-shot decision-making situation, excluding the complications that arise in sender-receiver (deception) games,

[2] See, among several others, Gneezy (2005); Sánchez-Pagés and Vorsatz (2007); Lundquist, Ellingsen, Gribbe, and Johannesson (2009); Erat and Gneezy (2012); Gibson, Tanner, and Wagner (2013); López-Pérez and Spiegelmann (2013).

[3] Accounting earnings management (henceforth referred to as "earnings management") occurs "when managers use judgment in financial reporting and in structuring transactions to alter financial reports to either mislead some stakeholders about the underlying economic performance of the company or to influence contractual outcomes that depend on reported accounting numbers" (Healy and Wahlen, 1999). In practice, there are also other motives for managing earnings, but we can exclude them in our experiment.

such as differences in the strategic sophistication of players (Cai and Wang, 2006; Wang, Spezio, and Camerer, 2010).

In this setting we study the effects of injunctive norms, describing specific kinds of behaviors that meet with real or perceived social approval or disapproval. Specifically, we randomly inform some participants of society's approval of earnings management, while others are informed of society's disapproval of earnings management.[4] As a second situational feature, we also vary the economic incentives in favor of lying.

We find that participants who are exposed to the situational norm approving (disapproving) of earnings management report the truth less (more) often than a control group does or than they themselves did before exposure to the situational norm. Economic incentives regarding truthfulness also exert influence in the expected direction.

Our main novel contribution consists of evidence demonstrating how the effects of situational norms vary across individuals. Using established scales (Tanner, Ryf, and Hanselmann, 2009), we measure the strength of an individual's commitment to honesty as a regard for truthfulness as a protected value. We primarily focus on "protected values: reactions to violations" (PRV), which measures the degree to which individuals experience affective reactions and emotional consequences when the value of honesty is violated or when the possibility of such a violation becomes salient (Tetlock, Kristel, Elson, Green, and Lerner, 2000).

We find that individuals with weak PRV conform more to situational social norms, whether those are dishonesty-approving or dishonesty-disapproving; individuals with strong PRV are more steadfast and less influenced by both types of norms. (We obtain these results as we control for the fact that participants with strong PRV are initially more likely to report the truth than those with weak PRV.) In other words, we find strong evidence of crowding out of intrinsic preferences for honesty by situational social norms.

[4] Injunctive situational norms entailing explicit penalties can be even stronger (though such penalties are not required for a social norm to be effective). In contrast to injunctive norms, descriptive norms simply describe the percentage of individuals choosing a particular behavior; see Cialdini, Reno, and Kallgren (1990).

Importantly, consistent with crowding out of intrinsic preferences by "good norms", intrinsic preferences provide a source of resistance against "bad norms" – the bright side of crowding out.

We also find less variability in responses to both situational norms (dishonesty-disapproving and dishonesty-approving) and economic incentives among those participants who express a pro-social concern (PSC) for certain stakeholders when acting dishonestly (though PRV explains more variability of behavior).

We then investigate various possible explanations for these results, thereby also checking the robustness of the findings. First, we document that experiment participation effects are unlikely to drive the results. Second, the results hold controlling for beliefs agents have regarding the consequences of their actions. Third, the findings also hold controlling for additional variables such as impression management motives.

Finally, we provide one possible explanation for *why* crowding out occurs in our experiment. Of three possible explanations – heterogeneous control aversion driving by a desire to self-signal, moral disengagement, and negative signals by situational norms – the first overall receives the most support. To clarify this idea, we propose a simple model of self-signaling, which is based on the premise that individuals are uncertain about their own true characters and that they interpret their actions as signals to themselves of their own preferences for truthfulness (Bem, 1972; Bodner and Prelec, 2002; Bénabou and Tirole, 2004, 2006). The model implies that, if resistance against situational norms is increasing in the true intrinsic preferences for truthfulness – as we find in the data – this supports a sorting condition that allows high "ethical" types to separate themselves from low types. Our additional tests – considering the role of demographics and other personal characteristics – further support the self-signaling hypothesis.

Our paper makes two contributions. First, it adds to the understanding of the effects of situational norms. In particular, this is among the first studies to document that the effects of situational norms vary with individuals' intrinsic costs of lying and in turn how the role of individual motivations varies with situational norms.[5] Second, with the conclusion regarding why crowding-out occurs in our study, this paper adds to the literature on self-signaling and self-image concerns in ethical behavior. We comment on this literature in Section 4 below.

The rest of this paper is organized as follows. Section 2 presents a theoretical framework and our hypotheses. Section 3 describes the experimental method and design. Section 4 presents the empirical results and their interpretation. Section 5 concludes.

2 Theoretical framework and hypotheses

2.1 Framework

2.1.1 Model

Consider a risk-neutral individual i who decides whether to tell the truth, $T = 1$, or to lie, $T = 0$. The agent takes two factors into account in this decision: his agent-specific costs of lying, and the extrinsic (situation-specific) consequences of lying.

On the one hand, individuals have intrinsic preferences for truthfulness. That is, they experience psychologically or morally driven costs of lying, $C_i = \theta_i V$, where V is the constant marginal value of truthfulness and θ_i is the agent ethical type (henceforth just "type"). Higher types value truthfulness more. C_i reflects true, intrinsic costs of lying.

On the other hand, there are extrinsic consequences of truthfulness, denoted by $EXCO$. We consider two kinds of extrinsic consequences: First, we take into account situational social norms regarding honesty. They cause **s***ituational* ***n****orm-driven* ***cost****s* ***o****f* ***s****tating the* ***t****ruth*, $SNCOST_s$, where s denotes the type of situational norm. A dishonesty-disapproving situational social norm operates against dishonesty; thus,

[5] Other work has examined the importance of situational context for ethical behavior. Situations that have been studied include ethical choices made at home versus in a public setting (Abeler, Becker, and Falk, 2014), a personal versus a market-focused context (Cappelen, Sørensen, and Tungodden, 2013), dictators' choices as a function of the context and of whether the recipient is informed about the context (Cappelen, Halvorsen, Sørensen, and Tungodden, 2013), competition-related situational norms (Rode, 2010; Fershtman, Gneezy, and List, 2012), and the role of voluntary versus imposed regulation and monitoring (Gino, Krupka, and Weber, 2013).

$SNCOST_{DISAPP} < 0$. Our analysis also applies to a dishonesty-approving situational social norm, implying $SNCOST_{APP} > 0$. A key element of these situational norm-driven costs of lying is that they are non-monetary: individuals may conform to norms because of threats of emotional penalties for norm violations and of disapproval, such as shame or accusations that the agent does not understand the unspoken "rules of the game" (Cialdini and Goldstein, 2004). In addition, conformist behavior in general may be instrumental in obtaining higher material rewards. We explain in the experimental analysis how we can, at least to some extent, distinguish between these two motivations associated with compliance to norms. Second, there are direct **e***conomic* ***c****osts* ***o****f* ***s****tating the* ***t****ruth*, $ECOST_e \geq 0$, where the index e denotes the $ECOST$ situation. In terms of the ultimate impact on utility, however, $ECOST_e$ and $SNCOST_s$ fulfill similar roles, and so for the purposes of the theoretical framework we subsume their impact under the term extrinsic consequences, $EXCO_{se}$.

The key ingredient of our analysis is that agents may differ in how they respond to $EXCO_{se}$. Let ρ_i indicate how much agent i resists extrinsic consequences. The consequences the agent perceives when telling the truth are $EXCO_{se}\left(1-\rho_i\right)$. Because wealth effects are unlikely in our experiment, we assume that all individuals have identical, constant marginal utility (set to unity for simplicity). In sum, the global utility function is

$$V_{ise}(T) = \begin{cases} -EXCO_{se}\left(1-\rho_i\right) & \text{if } T = 1 \\ -V\theta_i & \text{if } T = 0. \end{cases} \tag{1}$$

2.1.2 The role of ρ_i

We are in particular interested in whether the resistance parameter ρ_i is systematically associated with the agent's type θ_i, that is whether $\rho_i\left(\theta_i\right)$ is increasing or decreasing.

For simplicity, suppose that $\rho_i\left(\theta_i\right) = r\theta_i$. Depending on whether $r > 0$ or $r < 0$ or $r = 0$, resistance to extrinsic consequences is increasing in or decreasing in or independent of the ethical type. A natural benchmark is to posit that for some agents, truthfulness is a Kantian imperative, a “taboo value,” or a “sacred value,” meaning that the highest types, $\overline{\theta}$, are people who endorse protected values for truthfulness so completely that they resist all trade-offs (Fiske and Tetlock, 1997). Formally, when $r = 1/\overline{\theta}$, the highest type does not react at all to extrinsic consequences; when r is smaller (but still positive), all types respond at least to some extent to situational norms and economic incentives.

2.2 Hypotheses

Our empirical approach is to collect data on proxies for the (unobservable true) type and to then see whether these proxies are both correlated with truthtelling (as they should, if they proxy for the ethical type) *and* with resistance against situational norms and economic incentives.

Our primary proxy is a measure of the extent to which agents suffer negative emotional consequences when the value of honesty is or may be violated, called **p***rotected values for truthfulness associated with* ***r****eactions to* ***v****iolations of honesty* (PRV).[6]

We now state the model in terms of the parameters we will be able to identify empirically. For simplicity, we posit $\theta_i = tPRV_i$. Thus, $\rho_i = r\theta_i = rtPRV_i$.

From equation (1), combining and renaming coefficients, we can express the difference in utility between telling the truth and lying for individual i in economic situation e under situational social norm s as:

$$Y^*_{ise} = b_0 + b_P PRV_i + b_{EXCO} EXCO_{se} + b_{PEXCO} PRV^j_i EXCO_{se}. \tag{2}$$

Expanding equation (2) by separately considering $ECOST$ and $SNCOST$ and allowing dishonesty-approving ($SNCOST > 0$) and disapproving ($SNCOST < 0$) situational norms to have different effects,

[6] Below, we also report results for other proxies, including a measure of pro-social concern (PSC).

$$\begin{aligned} Y^*_{ise} &= b_0 + b_P PRV_i + b_{ASN} SNCOST_s 1_{\{SNCOST>0\}} + b_{DSN} SNCOST_s 1_{\{SNCOST<0\}} \\ &\quad + b_E ECOST_e \\ &\quad + b_{PASN} PRV_i SNCOST_s 1_{\{SNCOST>0\}} + b_{PDSN} PRV_i SNCOST_s 1_{\{SNCOST<0\}} \\ &\quad + b_{PE} PRV_i ECOST_e \\ &\quad + b_{EASN} ECOST_e SNCOST_s 1_{\{SNCOST>0\}} + b_{EDSN} ECOST_e SNCOST_s 1_{\{SNCOST<0\}}, \end{aligned} \tag{3}$$

where $1_{\{\}}$ is an indicator term, indicating whether a dishonesty-approving situational norm (abbreviated by ASN in the index) or a dishonesty-disapproving situational norm (DSN), respectively, is in place.

2.2.1 Direct impact of PRV

Naturally, we expect individuals with higher agent-specific costs of lying to perceive truthfulness as more attractive than lying. Thus, to the extent that PRV is a valid proxy for the type, we expect $b_P > 0$.

2.2.2 Hypothesis regarding direct responses to situational norms and economic incentives

A large literature (see the introduction) predicts that the pressure exerted by situational social norms leads individuals to conform with these norms. Situational norms are hypothesized to trigger an internal mechanism by which truthfulness becomes more or less attractive. Accordingly, we have:

Hypothesis CONFORM*: Truthfulness becomes less attractive under dishonesty-approving situational social norms and more attractive under dishonesty-disapproving situational social norms. Thus,* $b_{ASN} < 0$ *and* $b_{DSN} > 0$.

The alternative hypothesis, in line with Brehm's (1966) theory of reactance, suggests that agents tend to act in the opposite direction of what is suggested by the situational norm: $b_{ASN} > 0$ *and* $b_{DSN} < 0$.

Additionally, we expect that truthfulness becomes less attractive as the economic costs of truthtelling increase. Under our assumptions, $b_E = -1$.[7]

2.2.3 Hypothesis regarding heterogeneous responses to situational norms and economic incentives

Significant evidence suggests that economic incentives can induce "crowding out" of intrinsic preferences, though in several studies, in fact, "crowding in" occurred (see Bowles and Polanía-Reyes (2012) and Gneezy, Meier, and Rey-Biel (2011) for surveys). The reasons that Bowles and Polanía-Reyes (2012) list for why crowding out of intrinsic pro-social preferences by incentives supporting pro-social behavior can occur may also apply to the case of intrinsic preferences for truthfulness and dishonesty-disapproving situational norms ($SNCOST < 0$): (1) Dishonesty-disapproving situational norms may induce *control aversion* among those with strong intrinsic regard for truthfulness. (2) Another possibility is that these situational norms highlight the type of situation and may activate own payoff-maximizing modes of thought, thus inducing *moral disengagement.* (3) Also, addressees of these situational norms may understand them as a *signal* about the lack of trust by the principal (experimenter) in the agent (participant). Overall, under crowding out we expect $b_{PDSN} < 0$.

Importantly, our experiment also contains $SNCOST > 0$, and this allows us to test whether resistance occurs also on "the other side."[8] To the extent that symmetry holds, crowding out of intrinsic preferences for truthfulness by "good" norms actually has a bright side in that these intrinsic preferences will also weaken responses to "bad" norms.

[7] More generally, we expect b_E to be equal to minus the marginal utility of money. In addition, economic costs of truthfulness and the situational norm-driven costs may also affect preferences. If, for example, $ECOST$ is positively related to C_i, there is a countervailing effect. We cannot identify these effects within our study. Taking this possibility into account, we expect the attractiveness of truthfulness to not increase in economic costs of stating the truth, that is, $b_E < 0$. This is what we test in the empirical section.

[8] To see why a symmetric reaction is plausible, imagine that we measure "unethical" types and make lying the choice variable. Then, crowding-out would mean that more unethical individuals would respond less to the situational norm that approves of lying. Therefore, when truth-telling is the choice variable, the sign flips, and we thus expect ethical types to respond less to the dishonesty-approving norm.

In sum, we investigate the following hypothesis:

Hypothesis CROWD-OUT*: High PRV types resist both types of situational norms. Thus, $b_{PDSN} < 0$ and $b_{PASN} > 0$.*

We also expect resistance against economic incentives, that is $b_{PE} > 0$.[9]

As Bowles and Polanía-Reyes (2012) show, there are, however, also studies that find crowding in of intrinsic preferences by extrinsic incentives. Incentives may signal good news about the principal, and they may in fact lead to moral engagement. It is, thus, conceivable, that situational norms may crowd in intrinsic preferences, which would imply $b_{PDSN} > 0$, $b_{PASN} < 0$ and $b_{PE} < 0$. Finally, situational norms and economic incentives on the one hand and intrinsic preferences on the other hand may be separable, thus leading to $b_{PDSN} = 0$, $b_{PASN} = 0$ and $b_{PE} = 0$.

2.2.4 Hypothesis regarding interactions between economic incentives and situational norms

The role of economic costs may vary as the norms vary. For example, Fischer and Huddart (2008) derive a model in which social norms augment the effects of incentives. In our context, this model implies that a dishonesty-approving norm would complement the effects of economic incentives against truthfulness; a dishonesty-disapproving norm would work against $ECOST$. Thus:

Hypothesis ECOST-COMPL*: Economic costs of stating the truth and situational social norms are complements:* $b_{EASN} < 0$ and $b_{EDSN} > 0$*.*

Alternatively, economic incentives may weaken the effects of situational norms, perhaps by the same mechanisms by which they crowd out effects of intrinsic preferences. This would imply that $ECOST$ and situational social norms are substitutes, or $b_{EASN} > 0$ and $b_{EDSN} < 0$. Or, if preferences are separable in situational norms and economic incentives, we would have $b_{EASN} = b_{EDSN} = 0$.

[9] Gibson, Tanner, and Wagner (2013) document that agents with strong protected values react less to economic costs of stating the truth, but do not make the link to crowding-out and possible reasons, such as self-signaling, as the source of this resistance.

3 Experimental method

Our primary data come from an experiment presented in Gibson, Tanner, and Wagner (2013) (GTW). While that paper exclusively evaluated participant choices in the absence of an explicit manipulation of situational norms ("Phase 1"), in this study, we mainly evaluate participant choices in "Phase 2" of the experiment, which repeated the first decision but introduced explicit situational social norms. The full instructions are in the Supplementary Appendix.

3.1 Participants and procedure of the main experiment

A total of 261 participants took part in this online experiment. We recruited participants from undergraduate classes at the University of Zurich (Switzerland). 50 percent of the participants were economics and finance students, 40 percent psychology students, and 10 percent students from other fields. 42 percent were women, and 58 percent were men (distributed across the various fields).

All participants were informed at the outset that their choices would remain unknown to the experimenters. Most participants received payment one week after the experiment. For this purpose, each participant received, before the experiment, a code, based on which the experimenter prepared an envelope containing the earnings. Participants received the sealed envelopes by indicating their personal codes. They were first asked to respond to a few demographic questions and to read some basic instructions. They were informed that they would individually receive a payment, CHF 8 (about USD 9 at the time of the experiment), for their completed participation in the study, and an additional payment that depended on their decisions. After having demonstrated their understanding of the (unlabeled) tasks and of the rules of the experiment, the participants completed, in randomized orders, the four main parts of the experiment: 1) the truthtelling task (first without and then with situational norms), 2) the beliefs and manipulation check, 3) the effort task, and 4) the measurement of various controls and proxies for agent-specific costs of lying. Finally, all the participants were paid. For simplicity, we describe the procedure for one of the randomized orders of tasks.

1) The truthtelling task consisted of two phases, in each of which the participants were given five choices. Specifically, the task consisted of the following steps: choices in Phase 1, norm manipulation and choices in Phase 2.

In the truthtelling task within Phase 1, each participant was placed in the situation of a CEO who had to announce earnings per share for the previous quarter. The variable component of the CEOs' salaries depended on the earnings they announced. Participants were also informed that the market currently anticipated the announcement of 35 cents per share as earnings, but that the true earnings were 31 cents per share. The participants were told that they could announce earnings of 35 cents per share while remaining within legal accounting limits and that the decision would be solely theirs. Therefore, risk preferences of participants did not matter, as their choices were not based on the trade-off between the expected benefits and costs of committing a crime. They were also informed that they would be paid an amount based on the CEO compensation (according to their decisions). This additional experimental payoff would be converted into real money at the rate of CHF 100,000 = CHF 0.5. Importantly, participants earned less when choosing to tell the truth. Phase 1: The participants were then told they would have to announce their financial statements that day. The truthtelling task questionnaire follows, in paired questions:

Which earnings will you announce?

__ 31 cents per share – In this case, your compensation will be CHF 60,000 (CHF 0.30).

__ 35 cents per share – In this case, your compensation will be CHF 300,000 (CHF 1.50).

__ 31 cents per share – In this case, your compensation will be CHF 120,000 (CHF 0.60).

__ 35 cents per share – In this case, your compensation will be CHF 300,000 (CHF 1.50).

__ 31 cents per share – In this case, your compensation will be CHF 180,000 (CHF 0.90).

__ 35 cents per share – In this case, your compensation will be CHF 300,000 (CHF 1.50).

__ 31 cents per share – In this case, your compensation will be CHF 240,000 (CHF 1.20).

__ 35 cents per share – In this case, your compensation will be CHF 300,000 (CHF 1.50).

__ 31 cents per share – In this case, your compensation will be CHF 300,000 (CHF 1.50).

__ 35 cents per share – In this case, your compensation will be CHF 300,000 (CHF 1.50).

Phase 2, situational social norm manipulation: The novel feature of the present paper is that we analyze a manipulation that was introduced after Phase 1 and based on situational social norms. We faced the following trade-off in the experimental design. On the one hand, social norms affect observed actions; this would support revealing participants' choices in the experiment. On the other hand, we wished to avoid experimenter-demand effects. We chose an intermediate approach that is consistent with the overall design of the experiment as a decision-making situation in a concrete context. Participants knew that the market was observing their actions as CEOs; however, participants knew the experimenters could not learn their individual choices as experimental subjects. Specifically, the participants were given a page to read that stated that their respective firms would likely be confronted with a good investment opportunity the following year for the acquisition of another company. However, they would need the shareholders' approval for that project. At the shareholder meeting, they would have an opportunity to convince the shareholders of the soundness of this investment. These shareholders would be closely following the CEOs' earnings announcements as well as those of the competitors. With this information, which all participants received, we made it clear to participants that their actions as CEOs would be observed. Then, the participants were randomly assigned to one of the following three groups, which were not labeled for the participants:

I. CONTROL group

(No further information was provided beyond the common information.)

II. earnings-management-APPROVING situational social norm group:

"One evening, you are sitting with a friend of yours who is a financial analyst. He tells you that increasing reported earnings in order to meet market expectations meets with widespread societal approval."

III. earnings-management-DISAPPROVING situational social norm group:

"One evening, you are sitting with a friend of yours who is a financial analyst. He tells you that increasing reported earnings in order to meet market expectations meets with widespread societal disapproval."

After this interlude, all the participants were again provided with the same set of five options as in Phase 1, requiring them to choose to announce earnings of either 31 or 35 cents per share.

2) We then asked participants what they thought that the approval of the proposed acquisition project would depend on. We collected four belief measures. Participants could indicate, with yes/no answers, that they believed approval would depend on whether they had presented only high earnings, on how high their compensation was, on whether they were seen as competent, and on whether they had reported transparently in the past quarters. We also conducted a manipulation check. With the manipulation check, we verified that our participants perceived the announcement of 31 cents as the honest action that led to a personal loss, while the opposite was true of the announcement of 35 cents. We also measured participants' levels of pro-social concern (PSC).

3) Participants engaged in a simple effort (calculation) task. This task creates a time lag between the truthtelling task and the measurement of protected values for truthfulness. This task was also used to investigate (and alleviate) the concern that a positive or negative regard of participants for the wealth of the experimenters could explain behavior. The results of this analysis are reported in GTW.

4) We measured participants' levels of protected values for truthfulness and their tendencies towards impression management and self-deception.

The experiment lasted about 20 minutes, on average. The average total payment, received anonymously (see above) by each participant, was slightly less than CHF 30.5.

3.2 Variables of interest

TRUTHFUL CHOICE. This represents the dependent variable in the truthtelling task, coded as a binary variable that takes on the value of 1 if a participant chose to announce earnings of 31 cents (the honest option), while it takes on the value of 0 if a participant announced 35 cents (the dishonest option). TRUTHFUL CHOICE thus measures the extent of truthtelling, that is, the lack of earnings management.

SITUATIONAL SOCIAL NORM. This is a between-participants variation of *SNCOST*. The experiment did not offer continuous levels of *SNCOST*, but instead used three discrete levels. We define three dummies, making Phase 1 the omitted category in the regressions. CONTROL is equal to 1 for all observations from Phase 2 with no additional information, and to 0 otherwise. APPROVING is equal to 1 for all observations from Phase 2 with the situational social norm of approval of earnings management, and to 0 otherwise. DISAPPROVING is equal to 1 for all observations from Phase 2 with the situational social norm of disapproval of earnings management, and to 0 otherwise.

ECOST. This is a within-participants variation. Economic costs of truthfulness represent the amount of money a participant forfeited by announcing 31 cents. The *ECOST* variable takes on values from CHF 0 to CHF 1.20 (= 1.50 - 0.30), in increments of 30 cents.

AGENT TYPE. We use three proxies for ethical agent types: two proxies for a concern for and a commitment to truthfulness, and one proxy for pro-social concern. Two distinct subscales of protected values for truthfulness were developed in Tanner, Ryf, and Hanselmann (2009). PRV (reactions to violations) captures affective reactions to and emotional consequences of violations of honesty (Tetlock, Kristel, Elson, Green, and Lerner, 2000). PNT (no trade) captures the more cognitive notion of an individual's unwillingness to consider trade-offs that are based on an economic cost-benefit analysis when choosing between truthfulness and lying (Baron and Spranca, 1997). The details are available in Supplementary Appendix A.2.1. Both scales have high Cronbach's alpha (0.9 and 0.75, respectively). Both scales take on values between 0 (for an individual with no protected values) and 6 (for an individual with maximum protected values). We standardize the scales to have means of zero and standard deviations of unity (interdecile ranges: -1.25 to +1.24 and -1.25 to +1.47, respectively). The correlation between the two scales is 0.5, indicating that they are related but not identical. GTW showed that the combined average does a good job in explaining truthfulness; in the present paper, we consider the two scales separately because this allows us to shed additional light on distinct sources of truthfulness.

As our third proxy for agent-specific costs of lying, we consider pro-social concern (PSC). In an attempt to mirror reality, we are vague about the precise consequences for others and use participants' answers regarding the extent to which they believed that announcing 35 cents had negative consequences for some stakeholders or was manipulative. The details are available in Supplementary Appendix A.2.2. Participants who exhibited stronger pro-social concern score high on the resulting variable PSC. PSC is standardized to have a mean of zero and a standard deviation of unity (interdecile range: -1.57 to +1.16). PRV and PSC have a correlation of 0.39, and PNT and PSC have a correlation of 0.28.

DEMOGRAPHIC AND OTHER VARIABLES. Sex is equal to 1 for female participants and to 0 for male participants. Age is equal to each participant's age in completed years (interdecile range: 20 to 29 years). Economics is equal to 1 for economics and finance students and to 0 otherwise. Other is equal to 1 for students of other fields and to 0 otherwise. Psychology students are the omitted category in the regressions. We also collect data on whether participants had recently read newspaper articles regarding CEOs, whether they worked part-time, and whether they had investment experience.

4 Experimental results

4.1 Descriptive evidence

Table 1 allows a first look at the choices the participants made in the experiment. This table reveals substantial variation in responses throughout the within-participants and between-participants conditions that were established in the experiment. Of particular interest to our purpose are the choices made under the different situational norms. At the median cost level, around 31%-34% of participants told the truth in Phase 1 or in the CONTROL condition; by contrast, only 16% reported the truth under the earnings-management APPROVING norm, while 55% stated earnings truthfully under the DISAPPROVING norm. In aggregate, leaving aside the $ECOST = 0$ case, in 33% of cases, participants opted to suffer monetary losses relative to what they could have earned. However, even when there was no economic cost for truthfulness, 23% of the participants chose the earnings-management solution. These participants may have experienced negative costs of lying. Situational norms also had a significant impact at zero $ECOST$. In line with previous research, and informally supporting Hypothesis CONFORM, we thus find powerful direct effects of both types of situational social norms.

Table 1
Behavior across phases and costs of truthtelling

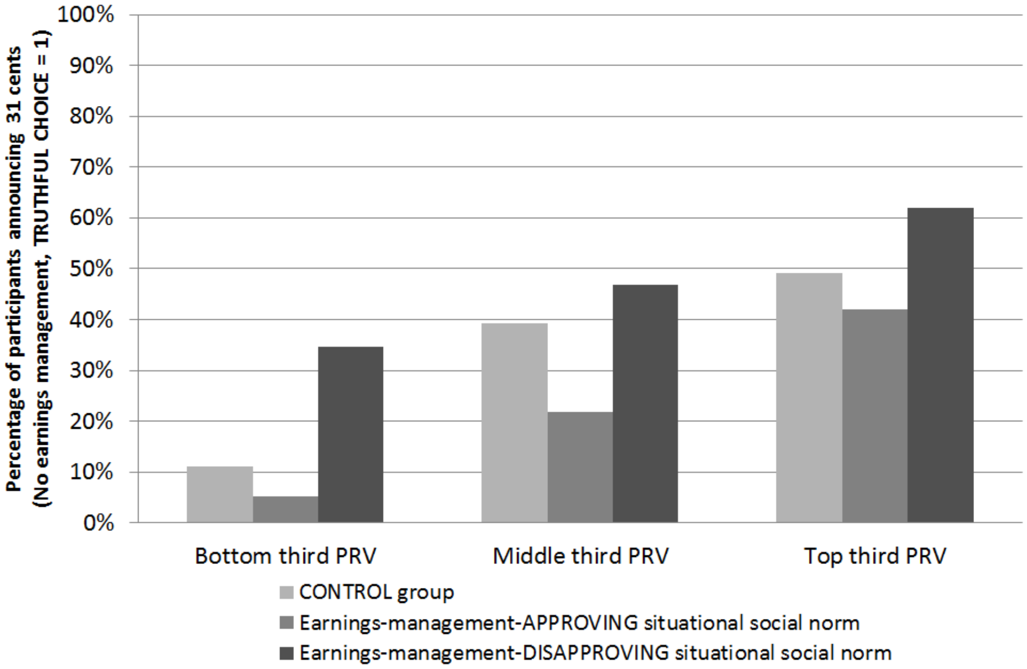
This table presents the percentages of participants announcing 31 cents of earnings per share (TRUTHFUL CHOICE = 1) across the various $ECOST$ conditions and phases (situational social norms conditions) of the experiment. Figure 1 (in the text) shows, for each of the three PRV (protected values for truthfulness, reactions to violations) terciles, the percentage of participants who announced 31 cents (TRUTHFUL CHOICE = 1) in Phase 2 of the experiment.

	Phase 1	**Phase 2 (earnings management norms)**			**All**
		APPROVING	**DISAPPROVING**	**CONTROL**	
ECOST	Percent of participants announcing 31 cents				
CHF 0	82.0%	63.1%	80.7%	71.9%	77.0%
CHF 0.3	52.1%	33.3%	65.9%	49.4%	51.0%
CHF 0.6	31.4%	15.5%	54.5%	33.7%	33.1%
CHF 0.9	23.0%	15.5%	35.2%	27.0%	24.5%
CHF 1.2	21.1%	14.3%	31.8%	23.6%	22.2%
Total	41.9%	28.3%	53.6%	41.1%	41.6%
Total except $ECOST = 0$	31.9%	19.6%	46.9%	33.4%	32.7%

Next, Figure 1 plots the levels of percentages of truthtellers across the three treatments of Phase 2 (over all *ECOST* situations), using the CONTROL group as the reference point. To provide some first insights into heterogeneous responses, the figure considers the behavior of participants in three PRV terciles separately.

Figure 1
Behavior across phases and costs of truthtelling

This figure shows, for each of the three PRV (protected values for truthfulness, reactions to violations) terciles, the percentage of participants who announced 31 cents (TRUTHFUL CHOICE = 1) in Phase 2 of the experiment. The figure uses all *ECOST* situations.



While Table 1 shows strong effects of the APPROVING norm in the sample on average, Figure 1 demonstrates that, among the top-third PRV group, behavior was quite stable in the face of the APPROVING norm. And while the DISAPPROVING norm more than doubled the number of truthtellers among the bottom-third PRV group, this norm had hardly any effect on behavior among the top-third PRV group. (These results are not simply due to the fact that the percentages of truthtellers cannot exceed 100%. In the top PRV tercile, the percentage of truthtellers in the CONTROL condition is in the 50% range, so that significant behavioral changes are, in principle, possible under the disapproving norm.) This descriptive evidence provides some first support for hypothesis CROWD-OUT.

4.2 Empirical model

We estimate a discrete-choice / random-utility model (King 1998; Wooldridge 2006). This model allows us to examine whether those with strong preferences for truthfulness respond less or more sensitively to situational norms and to economic incentives than those with weak preferences, in the hypothetical case that the probabilities of truthfulness are the same for participants with heterogeneous intrinsic preferences. In the following model statement, we abbreviate the situational norms binary indicator variables as APP, DISAPP, and CONT, respectively. (Phase 1 is the omitted category.) Starting from equation (3), assuming a stochastic error with a logistic distribution (independent of the explanatory variables), and positing that agents choose the action that provides them with the higher utility, one obtains the logit model, which is the main specification on which we focus. After relabeling coefficients,

$$\Pr\left(T_{ise}=1|\mathbf{X}\right)= \tag{4}$$

$$\Lambda\left[\begin{array}{c} \beta_0+\beta_P PRV_i+\beta_{ASN}APP_s+\beta_{DSN}DISAPP_s+\beta_{CSN}CONT_s+ \\ +\beta_E ECOST_e \\ +\beta_{PASN}PRV_iAPP_s+\beta_{PDSN}PRV_iDISAPP_s+\beta_{PCSN}PRV_iCONT_s \\ +\beta_{PE}PRV_iECOST_e \\ +\beta_{EASN}ECOST_eAPP_s+\beta_{EDSN}ECOST_eDISAPP_s+\beta_{ECSN}ECOST_eCONT_s \end{array}\right],$$

where $\Lambda\left(\bullet\right)$ is the logistic cumulative distribution function. If ε is normally distributed, one obtains the probit model. As is typical in econometric applications, the two models yield virtually identical inferences. The coefficient vector in Equation (4) is estimated by maximum likelihood. These coefficients are the implied estimates for the model parameters. We cluster standard errors on the individual level.

4.3 Regression results

4.3.1 Direct effects of situational norms and economic incentives

Table 2 shows that participants clearly responded strongly either to society's approval or to its disapproval of earnings management. These results support Hypothesis CONFORM, and they reject the idea of a uniform reactance against situational norms. The effects of social approval and disapproval of earnings management hold after controlling for PRV. Expressing as marginal effects the coefficients on the APPROVING and DISAPPROVING dummies implies that the approving norm made earnings management 15% more likely and that the disapproving norm made it 15% less likely.

The control group behaved about the same as in Phase 1. This suggests that the information that their behavior would be closely observed by shareholders did not by itself change participants' choices; it was the situational norm stating market participants' approval (or disapproval) of earnings management that triggered a behavioral change. (In untabulated results, we confirm that this result also holds if we restrict the sample to only those who, in Phase 2, were in the CONTROL group, comparing their behaviors in Phases 1 and 2).

As expected, the higher the economic incentives for earnings management were, the more likely participants were to manage earnings (see the negative coefficient on $ECOST$).

4.3.2 Intrinsic preferences and heterogeneous responses to situational norms and economic incentives

We now consider the relevance of differences in preferences across individuals. As is evident from column (1) of Table 2, PRV is strongly significantly positively associated with the perceived attractiveness of truthfulness. Demographic variables are not systematically related to truthfulness.

Our main results, regarding who responds the most to situational norms, referring to Hypothesis CROWD-OUT, are reported in columns (2) and (3) of Table 2.

Table 2
Heterogeneous responses to situational social norms

This table presents coefficients of logit regressions. The dependent variable is TRUTHFUL CHOICE, which is equal to 1 when a participant chose to announce 31 cents and equal to 0 otherwise. The construction of Protected valued for truthfulness - reactions to violations (PRV) is described in the text. PRV is standardized to have a mean of zero and a standard deviation of unity. Columns (1) and (2) use data from all *ECOST* situations. Column (3) considers situations where *ECOST* was strictly positive. All demographic controls are included; none have significant coefficients. T-statistics, obtained from robust standard errors clustered at the individual level, appear in parentheses below coefficient estimates. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

	(1)	(2)	(3) *ECOST*>0
PRV (protected values for truthfulness, reactions to violations)	0.69***	0.35***	0.37**
	(5.48)	(2.82)	(2.21)
*PRV * APPROVING*		0.39*	0.67**
		(1.95)	(2.29)
*PRV * DISAPPROVING*		-0.37*	-0.43**
		(-1.93)	(-2.23)
*PRV * CONTROL*		0.20	0.15
		(1.05)	(0.72)
*PRV * ECOST*		0.64***	0.55***
		(3.36)	(2.66)
APPROVING (situational norm, Phase 2)	-0.64***	-0.68***	-0.82***
	(-3.29)	(-3.54)	(-3.12)
DISAPPROVING (situational norm, Phase 2)	0.59***	0.64***	0.77***
	(3.37)	(3.66)	(4.46)
CONTROL group (Phase 2)	-0.16	-0.22	-0.11
	(-0.88)	(-1.21)	(-0.53)
ECOST (cost of no earnings management)	-2.36***	-2.51***	-1.89***
	(-13.97)	(-12.87)	(-8.62)
Demographic controls (all insignificant)	Yes	Yes	Yes
Constant	1.59**	1.46*	0.45
	(1.97)	(1.87)	(0.56)
Observations	2,610	2,610	2,088
Pseudo R^2	0.208	0.219	0.161
Pseudo Log Likelihood	-1403	-1384	-1107
Likelihood-ratio test statistic (χ^2, p-value)	738 (<0.01)	776 (<0.01)	426 (<0.01)
Wald test statistic (χ^2, p-value)	230 (<0.01)	237 (<0.01)	121 (<0.01)

Column (2) uses all *ECOST* situations, while column (3) only considers situations with positive *ECOST*.

We find a negative coefficient on the interaction PRV * DISAPPROVING and a positive coefficient on the interaction term PRV * APPROVING. In other words, those with stronger PRV responded less to the dishonesty-disapproving situational norm than those with weaker PRV. In the presence of the dishonesty-disapproving situational norm, these intrinsic preferences were less important in guiding individuals towards truthfulness than in the absence of that norm. Thus, we find evidence of crowding-out.

Interestingly, we also find that those with strong protected values were steadfast in the face of a dishonesty-approving situational norm. In other words, the effect of intrinsic preferences for truthfulness due to emotional reactions to violations of honesty is particularly pronounced when truthfulness is socially devalued. Recall that we have standardized the measures of intrinsic costs of lying. Because of the standardization, the coefficients shown for APPROVING represent the effects for a person of average PRV, in which case the standardized PRV score is zero, so that the interaction term with PRV cancels out. A person with PRV one standard deviation above the mean reacted only about half as strongly to an approving norm (-0.68 + 0.39 = -0.29 instead of -0.68) as the mean participant; in the range of positive *ECOST*, the effect was stronger (-0.82 + 0.67 = -0.15).[10] PRV also induces significant resistance against economic costs.

As a complementary approach, to test the robustness of the findings and to investigate potential non-linearities, in Table 3 we consider the impact of situational norms separately for different quantiles of PRV, that is, non-parametrically. The maximum PRV of participants in the first quartile is -0.71; the minimum PRV of participants in the fourth quartile is +0.71.

[10] As discussed by Bowles and Polanía-Reyes (2012), non-separability can be categorical or marginal; that is, there may be a distinct effect of introducing any positive amount of *ECOST*. The findings here also suggest that the mere presence of monetary stakes has categorical effects on the role of intrinsic preferences and on their interplay with situational norms.

4.3.3 Interactions between situational norms and economic incentives

Our experiment also allows us to test Hypothesis ECOST-COMPL. To study how economic incentives interact with the situational norm, regressions (5) to (8) in Panel B of Table 3 add the interactions between *ECOST* and situational norms as an explanatory variable. We find that the coefficient on *ECOST*APPROVING* is strongly negative for participants with low PRV. In other words, for the typical person who is not motivated by moral preferences, Hypothesis ECOST-COMPL is a good description of reality. This corresponds to the prediction of the model in Fischer and Huddart (2008). By contrast, for those with strong PRV, the *ECOST*APPROVING* interaction term is in fact positive, though insignificant. The difference between Q1 and Q4 is highly significant, consistent with Hypothesis CROWD-OUT. The results regarding steadfastness with respect to the disapproving norm are somewhat weaker, though they trend in the same direction.

Columns (1) to (4) in Panel A show that participants with higher PRV values were generally more impervious to both norms as well as to economic incentives, in accordance with Hypothesis CROWD-OUT. In the case of the disapproving norm, the coefficients between the quantiles are not monotonic; they are, however, overall decreasing from the first to the fourth quartiles. This analysis highlights that the participants with protected values significantly below the median were those who drove the strong response to situational norms.[11]

[11] Naturally, the regressions also imply that the marginal effect on the probability of truthtelling of approving and disapproving social norms was greatest for participants whose protected values approximated the median (not shown). Intuitively, for those who were strongly opportunistically inclined, social norms regarding earnings management did not have measurable behavioral effects, either because an approving norm encountered people who were already lying or because a disapproving norm failed to dislodge participants who were initially unmotivated to consider truthfulness as a viable option.

Table 3

Differential resistance to situational social norms–non-parametric analysis

This table presents coefficients of logit regressions. The dependent variable is TRUTHFUL CHOICE, which is equal to 1 when a participant chose to announce 31 cents and equal to 0 otherwise. The regressions are calculated separately for the participants in the quantiles of PRV described at the tops of the respective columns. The explanatory variables are defined in the text and in the notes to Table II. T-statistics, obtained from robust standard errors clustered at the individual level, appear in parentheses below coefficient estimates. All demographic controls are included; none have significant coefficients. The final column shows differences between coefficient estimates of interest for the top and bottom quantiles relevant to the respective regression sets are shown, for the variables listed in each row. The z-statistics for the significance of these differences are in parentheses. As we have independent samples, these statistics are computed as $(\beta_i - \beta_j) / \left(\sqrt{se(\beta_i)^2 + se(\beta_j)^2}\right)$ for two quantiles i, j. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

	(1)	(2)	(3)	(4)	
Panel A	Bottom quarter	Q2	Q3	Top quarter	Difference
	PRV	PRV	PRV	PRV	Q4 - Q1
APPROVING (situational norm, Phase 2)	-1.23***	-0.80**	-0.72**	0.08	1.31**
	(-3.44)	(-2.03)	(-2.19)	(0.18)	(2.24)
DISAPPROVING (situational norm, Phase 2)	1.30***	0.37	0.73**	0.26	-1.04**
	(3.42)	(1.05)	(2.19)	(0.83)	(2.12)
CONTROL group (Phase 2)	-1.13***	-0.06	0.30	-0.35	0.78*
	(-2.99)	(-0.13)	(0.94)	(-1.26)	(1.65)
ECOST (cost of no earnings management)	-3.52***	-2.52***	-2.44***	-1.63***	1.89***
	(-6.41)	(-7.08)	(-7.33)	(-6.61)	(3.13)
Demographic control variables	Yes	Yes	Yes	Yes	
Constant	1.36	0.21	1.42	1.53	
	(0.71)	(0.12)	(1.15)	(0.93)	
Observations	670	640	710	590	
Pseudo R^2	0.3	0.18	0.17	0.1	
Pseudo Log Likelihood	-361.5	-481.3	-481.3	-489.8	
Likelihood-ratio test statistic (χ^2, p-value)	214 (<0.01)	150 (<0.01)	172 (<0.01)	80 (<0.01)	
Wald test statistic (χ^2, p-value)	96 (<0.01)	60 (<0.01)	57 (<0.01)	51 (<0.01)	

	(5)	(6)	(7)	(8)	
Panel B	Bottom quarter	Q2	Q3	Top quarter	Difference
	PRV	PRV	PRV	PRV	Q4 - Q1
*(ECOST*APPROVING)*	-4.40**	-0.33	-0.05	1.07	5.48***
	(-2.23)	(-0.33)	(-0.07)	(1.61)	(2.64)
*(ECOST*DISAPPROVING)*	1.63**	0.29	0.71	0.43	-1.20
	(2.22)	(0.53)	(0.94)	(1.22)	(1.47)
*(ECOST*CONTROL)*	-3.74**	0.94*	0.60	0.13	3.86**
	(-2.05)	(1.70)	(1.37)	(0.33)	(2.07)
APPROVING (situational norm, Phase 2)	-0.70	-0.69	-0.72*	-0.60	0.04
	(-1.54)	(-1.63)	(-1.95)	(-1.05)	(0.05)
DISAPPROVING (situational norm, Phase 2)	0.45	0.21	0.27	-0.03	-0.48
	(1.09)	(0.47)	(0.44)	(-0.10)	(0.94)
CONTROL group (Phase 2)	-0.62	-0.54	-0.06	-0.44	0.18
	(-1.38)	(-1.28)	(-0.16)	(-1.44)	(0.32)
ECOST (cost of no earnings management)	-3.76***	-2.72***	-2.65***	-1.86***	1.90***
	(-5.46)	(-6.81)	(-7.40)	(-6.90)	(2.56)
Demographic control variables	Yes	Yes	Yes	Yes	
Constant	1.48	0.33	1.54	1.69	
	(0.77)	(0.18)	(1.25)	(1.02)	
Observations	670	640	710	590	
Pseudo R^2	0.32	0.18	0.18	0.1	
Pseudo Log Likelihood	-353.2	-479.9	-479.9	-488.6	
Likelihood-ratio test statistic (χ^2, p-value)	230 (<0.01)	153 (<0.01)	174 (<0.01)	82 (<0.01)	
Wald test statistic (χ^2, p-value)	94 (<0.01)	61 (<0.01)	62 (<0.01)	60 (<0.01)	

4.4 Drilling deeper

In this section, we document that pro-social concern is also related to resistance against norms (Section 4.4.1). Then, we find that controlling for beliefs of participants regarding the determinants of project approval by shareholders (Section 4.4.2) does not affect the results. Moreover, the results are robust to controlling for impression management motives (Section 4.4.3) and marginal utility (Section 4.4.4). We also show that the experiment itself is unlikely to have had an effect on individuals' protected values that would distort our results (Section 4.4.5). Section 4.4.6 contains some robustness checks.

4.4.1 Pro-social concern

An alternative proxy for ethical preferences is pro-social concern (PSC). Significant evidence exists on the role of this motivation (Fehr and Fischbacher, 2002, 2003). In the context of charitable donations, DellaVigna, List, and Malmendier (2012) propose a structural model whose estimates indicate that both altruism and social pressure are important determinants of giving. Relatively little is known, however, about how pro-social concern interacts with situational norms. Regression (1) of Table 4 shows that PSC's results parallel those of PRV. Thus, crowding out, rather than crowding in of pro-social concern by situational social norms is documented in our experiment. However, once we include PRV into the regression, the significance of PSC in explaining heterogeneity in responses is diminished.[12] This is consistent with the design of the experiment, in which strategic and pro-social motivations are less likely to play a role.

[12] In the results presented here, we orthogonalize PRV and PSC, but similar results also hold if we include the main measures.

Table 4
Additional results: Pro-social concern, beliefs, impression management

This table presents coefficients of logit regressions. The dependent variable is TRUTHFUL CHOICE, which is equal to 1 when a participant chose to announce 31 cents and equal to 0 otherwise. PRV and PSC are standardized to have a mean of zero and a standard deviation of unity, and are orthogonalized. Participants could indicate, with yes/no answers, that they believed project approval would depend on whether they had presented only high earnings (Belief-earnings), on how high their compensation was (Belief-compensation), on whether they were seen as competent (Belief-competence), and on whether they had reported transparently in the past quarters (Belief-transparency). The regressions consider situations where *ECOST* was strictly positive. T-statistics, obtained from robust standard errors clustered at the individual level, appear in parentheses below coefficient estimates. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

	(1)	(2)	(3)	(4)
PRV (protected values for truthfulness)		0.40**	0.32*	0.39**
		(2.36)	(1.73)	(2.29)
*PRV * APPROVING*		0.64**	0.94***	0.50*
		(2.17)	(2.63)	(1.65)
*PRV * DISAPPROVING*		-0.40**	-0.55***	-0.40**
		(-2.11)	(-2.86)	(-2.04)
*PRV * CONTROL*		0.10	0.15	0.18
		(0.46)	(0.68)	(0.81)
*PRV * ECOST*		0.53***	0.66***	0.51**
		(2.59)	(2.81)	(2.44)
PSC (pro-social concern)	0.18	0.09		
	(1.13)	(0.60)		
*PSC * APPROVING*	0.62*	0.09		
	(1.80)	(0.25)		
*PSC * DISAPPROVING*	-0.48**	-0.29		
	(-2.41)	(-1.59)		
*PSC * CONTROL*	0.38*	0.38**		
	(1.86)	(2.02)		
*PSC * ECOST*	0.74***	0.46*		
	(3.06)	(1.93)		
Belief-earnings (1: yes, 0: no)			-1.21***	
			(-5.06)	
Belief-compensation (1: yes, 0: no)			-0.33	
			(-0.97)	
Belief-competence (1: yes, 0: no)			-0.04	
			(-0.08)	
Belief-transparency (1: yes, 0: no)			1.24***	
			(2.83)	
IMPRESS				-0.00
				(-0.00)
*IMPRESS * APPROVING*				0.78***
				(2.97)
*IMPRESS * DISAPPROVING*				-0.19
				(-0.99)
*IMPRESS * CONTROL*				0.07
				(0.37)
*IMPRESS * ECOST*				0.31*
				(1.73)
ECOST (cost of no earnings management)	-1.92***	-2.02***	-2.13***	-1.95***
	(-8.84)	(-8.98)	(-8.77)	(-8.76)
APPROVING (situational norm, Phase 2)	-0.88***	-0.86***	-0.81***	-0.96***
	(-3.45)	(-3.50)	(-3.01)	(-3.27)
DISAPPROVING (situational norm, Phase 2)	0.76***	0.80***	0.72***	0.79***
	(4.43)	(4.65)	(4.04)	(4.55)
CONTROL group (Phase 2)	-0.06	-0.12	-0.07	-0.15
	(-0.31)	(-0.58)	(-0.30)	(-0.72)
Demographic control variables	Yes	Yes	Yes	Yes
Constant	0.57	0.61	0.56	0.61
	(0.76)	(0.78)	(0.50)	(0.76)
Observations	2,088	2,088	2,088	2,088
Pseudo R^2	0.150	0.186	0.238	0.179
Pseudo Log Likelihood	-1122	-1074	-1006	-1084
Likelihood-ratio test statistic (χ^2, p-value)	396 (<0.01)	492 (<0.01)	628 (<0.01)	472 (<0.01)
Wald test statistic (χ^2, p-value)	130.8 (<0.01)	154.2 (<0.01)	144.3 (<0.01)	119.0 (<0.01)

4.4.2 Beliefs

In column (3) of Table 4, we include as additional explanatory variables the beliefs that participants held regarding the determinants of project approval by shareholders. We find that participants who believed that project approval after Phase 2 would depend on whether they had always announced high earnings were more likely to report 35 cents of earnings, whereas those who believed that project approval depended on transparency were more likely to report the true earnings. Perceived competence played no role for decisions. Importantly, our main findings remain robust.

4.4.3 Impression management

It is possible that participants were signaling for social image reasons (under the assumption that they believed that the experimenter positively valued honesty), as in Bernheim (1994)). Thus, participants might have had an interest in pleasing the experimenter by appearing honest and non-greedy (Fischbacher and Heusi, 2013). The anonymity, relative to the experimenters, makes it unlikely that this was a major factor in our experiment.However, some researchers have argued that social esteem may, in fact, play a role even in anonymous settings (Ellingsen and Johannesson, 2008). We can also at least to some extent control for this factor by including a measure of individuals' tendencies towards impression management, IMPRESS (see Supplementary Appendix A.2.3). While there is some evidence that IMPRESS interacted with situational norms, Column (4) in Table 4 shows that our results regarding Hypotheses CONFORM and CROWD-OUT are robust even when controlling for this factor.[13]

[13] If anything, those with a stronger tendency to impress others by adhering to the social norm responded less to the earnings-management-disapproving norm; this is against what one would expect if impression management was driving responses to social norms. The interaction term is insignificant, though. There does appear, however, to be significantly stronger resistance to the approving norm among participants who are more prone to impression management.

4.4.4 Marginal utility of money

It is possible that the experiment or the situational norms manipulation differentially activated payoff-maximizing modes of thought. In an attempt to address this possibility, we control for a self-reported measure of marginal utility, MU (see Supplementary Appendix A.2.4 for details). In results available on request, we find that (1) MU is not correlated with PRV, (2) MU does not differ across the situational norms manipulations, and (3) MU is insignificant when included in the regressions (and all other results remain).

4.4.5 Experiment-participation effects

A potential concern as regards our findings is that the experiment may affect participants' answers on the protected values survey. Either the experiment itself or the situational norms manipulation may conceivably have played a role.

Collectively, three empirical observations mitigate this concern. First, we conducted a separate survey with 123 economics students who did not participate in any experiment (the non-participants). We only measured the protected values of the students in this sample, and we did not involve them in any of the choice tasks. The means of the two PRV distributions of economics-student participants and non-participants are not statistically different (p-value of 0.24). Before rescaling, the average PRV of economics-student participants (non-participants) was 3.87 (3.68) while the 10th, 25th, 75th, and 90th percentiles were 2.6 (2.2), 3 (2.8), 4.8 (4.6), and 5.2 (5.6), respectively. Kolmogorov-Smirnov tests do not reject the hypothesis that the two PRV distributions are the same (p-value of 0.4).

Second, PRV is measured with a time lag after an interim (effort) task. This mitigates the concern that participants tried to answer the protected values survey consistently with their choices on the truthtelling task.

Third, there is no statistically significant difference in either protected values scale between those who went through the approving-norm treatment and those who went through the disapproving-norm treatment. And given that we control for situational norms in the regressions, any such effect would be accounted for.

4.4.6 Robustness

We have considered many variations of the analysis. Results available upon request show that models based on random effects logit regressions yield results similar to those from before. Additional results include the following. Comparing the behaviors of the participants in the cost and in the timing randomizations, we find no statistically significant differences. Using a logarithmic transformation of PRV slightly strengthens the results. Defining PSC to include also the extent to which participants regard announcing 35 cents as corresponding to a personal gain (thus presumably corresponding to a loss for somebody else) and as short-term behavior yields similar results.

4.5 Interpretation: Why does crowding out occur?

We have established robust evidence that situational norms crowd out intrinsic preferences for truthfulness, as proxied for by PRV. Fortunately, this resistance, of high-PRV individuals, against "good norms" also translates into a stronger resistance against "bad" norms. But what may be behind this resistance?

It is not possible to definitively rule out any of the three channels, listed in Section 2.2.3, that Bowles and Polanía-Reyes (2012) posit for why non-separability of extrinsic incentives (and, by analogy, situational norms) and intrinsic preferences may occur: (1) control aversion, (2) moral disengagement, and (3) incentives (in our case: situational norms) as a negative signal. However, given the anonymity of the experiment, and given the lack of evidence that experiment participation effects may have played a role (see Section 4.4.5), we believe that crowding-out in our experiment is unlikely to be due to the situational norms having served as a signal of the belief of a principal towards an agent (channel (3)). Also, given that the situational norms manipulation did not change the payoff structure of the situation, and given the findings regarding the insignificant impact of the norms manipulation on self-reported marginal utility (see Section 4.4.4), we regard it as rather unlikely that moral disengagement (channel (2)) is a primary driving force of crowding-out in our experiment.

This leaves channel (1), heightened control aversion by those with strong PRV, as a possible explanation for our results. One approach to understand control aversion from an economic point of view can be found in a self-signaling framework.[14] This model class fits our context well because in our experiment, too, there is no direct observer of the participants' actions.

In Supplementary Appendix A.1 we propose a simple model in this spirit. An agent, when reflecting on his prior choices in order to infer his own true (ethical) types from his action, understands that his actions depend both directly on his intrinsic preferences as well as on how strongly he responds to situational norms (and economic incentives). We prove that high types can, by resisting situational norms (and economic incentives) more than others, credibly self-signal their identities as truthful or pro-socially oriented individuals, respectively. Formally, a positive correlation of types and resistance ($r > 0$ in the notation of Section 2.1.2) supports a self-signaling equilibrium.

Thus, a self-signaling model is consistent with the evidence obtained so far. Surely, it is implausible that *only* self-signaling drives behavior, and it is unlikely that one could reject all other conceivable explanations for truthfulness. What we can consider, however, is whether the self-signaling model has additional implications beyond what we have observed so far. This is indeed the case. The two additional implications have to do

[14] See Bodner and Prelec (2002) and Bénabou and Tirole (2004) for the first self-signaling models. Bénabou and Tirole (2006) formalize the idea that, in the presence of extrinsic incentives, one's action may be a less valuable signal of one's own true preferences. Self-signaling models provide an interpretation of what happens when individuals engage in self-regulatory processes by which they control their behaviors so as to live up to their own intrinsic moral standards (Bandura, 1986; Aquino and Reed, 2002). Self-signaling can help build an identity, which can be an "asset" (Bénabou and Tirole, 2011). Mazar, Amir, and Ariely (2008) argue that people may, at least to some extent, behave truthfully because they have a desire to maintain their self-concepts as honest persons (see also Fischbacher and Heusi (2013)). While self-deception is at the core of the Mazar, Amir, and Ariely (2008) analysis, the economic framework of self-signaling builds on Bayesian signaling. See Mijovic-Prelec and Prelec (2010) for a model of self-deception as a self-signaling. In results available on request, we find that the previous results are robust when tendencies for self-deception (see Supplementary Appendix A.2.3) are accounted for.

with the general point that self-signaling can only occur regarding agent characteristics for which the action in question actually is an informative signal.

First, demographics should not matter, because no self-signaling can take place with respect to known individual characteristics such as gender. As is documented in the Supplementary Appendix, Table A.1, men and women responded similarly to situational norms,[15] as did younger and older participants. Also, the extent to which participants had previously read newspaper articles about CEOs did not interact with situational norms in determining their truthful choices. Similarly, economics students did not respond to situational norms differently from psychology students, nor did participants with investment experience respond differently from those without investment experience. These latter results also suggest that the influence of the situational social norms and, in particular, their interactions with intrinsic costs of lying were not limited to participants who were already familiar with the experiment's subject matter. Demographics generally do not explain the responsiveness to economic incentives, either. (Some regressions suggest that, while women and non-investors told the truth more on average, they responded more strongly to economic incentives but this evidence is not robust across specifications.)

Second, recall that PNT measures the cognitive notion that individuals differ in the extent to which they regard truthfulness as priceless and beyond the scope of an economic cost-benefit analysis (Baron and Spranca, 1997). If an agent wishes to signal to himself that he is a non-consequentialist, reporting the truth is an informative signal especially when money is at stake. Thus, we expect $b_{AE} > 0$ for PNT. By contrast, if situational norms do not bring about instrumental benefits – as is the case in our experiment – cost-benefit considerations are not directly applicable to begin with. Thus, after observing his own action, the agent's posterior type estimate is equal to the prior, that is, the agent cannot draw any inferences regarding his identity as a non-consequentialist from his responses to such situational norms. Therefore, if it is self-signaling that drives people

[15] Our finding that preferences for truthfulness of both women and men are stable across situational norms is of interest, as other work suggests that women's social preferences are more malleable by context than men's (Croson and Gneezy, 2009).

towards truthfulness, we would expect $b_{AASN} = b_{ADSN} = 0$ for PNT. In results available on request, we find evidence in line with these predictions.

Overall, the very rich set of evidence obtained in this experiment can be explained with the self-signaling model, which provides an economic interpretation of (part of) the concept of control aversion. Seen from this perspective, what our paper adds to the existing literature is the link between the self-signaling framework and individuals' heterogeneous responses to situational social norms.[16]

5 Discussion and conclusion

We conduct a simple, anonymous, and non-strategic earnings-management experiment using actual monetary incentives to lie. We introduce injunctive situational social norms, allowing for both the approval and the disapproval of earnings management by society. The central contribution of the paper consists in linking reactions to situational norms to individual-level characteristics.

First, regarding the effects of situational norms, anecdotal evidence suggests that conformity to social norms is not uniform. Even under extreme circumstances, some individuals resist social norms. For example, some individuals risked their lives to save others from persecution by the Nazis, even though the most prevalent social norm pointed towards approval or at least tacit acceptance of such persecution. However, this is one of the first papers to study systematically the potential differences in individuals' responses to situational norms. Our results support the idea that individuals' responses to social norms are heterogeneous. More specifically, our paper documents that part of the het-

[16] Existing evidence on the role of self-signaling is inconclusive. Grossman and van der Weele (2013) provide evidence consistent with self-signaling by considering agent's decisions to avoid collecting information about a possible negative social impact of their decisions (see also Dana, Weber, and Kuang (2007)). The experiment of van der Weele and von Siemens (2014) does not provide support for a basic aspect of self-signaling, namely, that knowing that one will, in the future, be reminded of one's present charitable actions induces more charitable actions today. The experimental evidence in Grossman (2012) offers stronger support of social signaling than Bayesian self-signaling as drivers of acts of giving.

erogeneity can stem from individuals' intrinsic preferences.[17] Our results suggest that contagion effects (Gino, Ayal, and Ariely, 2009) can be mitigated by hiring agents who feel strongly about violations of honesty or who have strong pro-social concern. They also suggest that, when corporate and/or social policies like the ones mentioned in the introduction rely on situational norms, heterogeneity in the responses and potential resistance by some agents to such norms are to be taken into account.

Second, regarding the sources of truthfulness, it is worth noting that our results do not reject the possibility that true, "deep" and fully non-consequentialist preferences for truthfulness may drive honest behavior. Moreover, in reality, other factors such as repeated interaction and more explicit punishment for lying are also likely to drive individuals towards truthfulness. However, if one wants to isolate a single theoretical framework that can explain all of the evidence stemming from our experiment, we show that a model of self-signaling in the spirit of Bodner and Prelec (2002) and Bénabou and Tirole (2006, 2011) seems appropriate.

The main point of the paper is that situational pressure and related norm-based explanations for ethical (and unethical) behavior on the one hand and individual-level explanations such as self-signaling on the other hand are not separable but interact. While our findings suggest that situational norms disapproving of dishonesty crowd out intrinsic preferences for truthfulness, the good news is that when the situational norm approves of dishonesty, those with a strong commitment to honesty will resist that "bad" norm. This reveals the bright side of crowding-out.

[17] As such, our paper complements work by Kimbrough and Vostroknutov (2014) who find heterogeneity in responsiveness to social norms, but do not relate that heterogeneity to intrinsic preferences and do not consider the relationship between norm resistance and self-signaling. Formally, consider the following theoretical papers, using the notation in the respective studies. Our results suggest that the sensitivity parameter with respect to the social and personal norms in Fischer and Huddart (2008), α_i, is a function of the agent's personal norm, A_i. Similarly, in Burks and Krupka (2012), the parameter that determines how strongly an individual adheres to the group ethical norm, γ_i, would be a function of the personal norm function N_i. The function that characterizes externalities to the agent and to others in Huck, Kübler, and Weibull (2012), g_i, would be a function of intrinsic social preferences, which are not separately modeled in their paper. In Sliwka (2007), there are some absolutely steadfast individuals (who are either selfish or ethical/fair) and some absolutely conformist individuals. Although our work pertains to the case of truthfulness only, our results suggest that the degree of conformity regarding a social norm for fairness might be related to intrinsic preferences for fairness.

References

Abeler, Johannes, Anke Becker, and Armin Falk, 2014, Representative evidence on lying costs, *Journal of Public Economics* 113, 96–104.

Ajzen, Icek, and Martin Fishbein, 1980, *Understanding Attitudes and Predicting Social Behavior* (Prentice-Hall: Englewood Cliffs).

Andreoni, James, and B. Douglas Bernheim, 2009, Social image and the 50-50 norm: A theoretical and experimental analysis of audience effects, *Econometrica* 77, 1607–1636.

Aquino, Karl, and Americus Reed, 2002, The self-importance of moral identity, *Journal of Personality and Social Psychology* 83, 1423–1440.

Asch, Solomon E., 1956, Studies of independence and conformity: a minority of one against a unanimous majority, *Psychological Monographs* 70, 1–70.

Bandura, Albert, 1986, *Social foundations of thought and action* (Prentice-Hall: Englewood Cliffs).

Baron, Jonathan, and Mark Spranca, 1997, Protected values, *Organizational Behavior and Human Decision Processes* 70, 1–16.

Bem, Daryl J., 1972, *Self-Perception Theory*vol. 6 . pp. 1–62 (McGraw-Hill).

Bénabou, Roland, and Jean Tirole, 2004, Willpower and personal rules, *Journal of Political Economy* 112, 848–886.

——— , 2006, Incentives and prosocial behavior, *American Economic Review* 96, 1652–1678.

——— , 2011, Identity, morals, and taboos: Beliefs as assets, *Quarterly Journal of Economics* 126, 805–855.

Bernheim, B. Douglas, 1994, A theory of conformity, *Journal of Political Economy* 102, 841–877.

Bodner, Ronit, and Drazen Prelec, 2002, Self-signaling and diagnostic utility in everyday decision making, in Isabelle Brocas, and Juan D. Carrillo, ed.: *Collected essays in psychology and economics* . pp. 105–126 (Oxford University Press: Oxford).

Bowles, Samuel, and Sandra Polanía-Reyes, 2012, Economic incentives and social preferences: Substitutes or complements?, *Journal of Economic Literature* 50, 368–425.

Brehm, Jack, 1966, *A theory of psychological reactance* (Academic Press: New York).

Burks, Stephen V., and Erin L. Krupka, 2012, A multimethod approach to identifying norms and normative expectations within a corporate hierarchy: Evidence from the financial services industry, *Management Science* 58, 203–217.

Cai, Hongbin, and Joseph Tao-Yi Wang, 2006, Overcommunication in strategic information transmission games, *Games and Economic Behavior* 56, 7–36.

Cappelen, Alexander W., Trond Halvorsen, Erik Ø. Sørensen, and Bertil Tungodden, 2013, Face-saving or fair-minded: What motivates moral behavior?, *Working paper, NHH.*

Cappelen, Alexander W., Erik Ø. Sørensen, and Bertil Tungodden, 2013, When do we lie?, *Journal of Economic Behavior and Organization* 93, 258–265.

Cialdini, Robert B., and Noah J. Goldstein, 2004, Social influence: Compliance and conformity, *American Review of Psychology* 55, 591–621.

Cialdini, Robert B., Carl A. Kallgren, and Raymond R. Reno, 1991, A focus theory of normative conduct, *Advances in Experimental Social Psychology* 24, 201–234.

Cialdini, Robert B., Raymond R. Reno, and Carl A. Kallgren, 1990, A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places, *Journal of Personality and Social Psychology* 58, 1015–1026.

Cohn, Alain, Ernst Fehr, and Michel Maréchal, 2014, Business culture and dishonesty in the banking industry, *Nature* 516, 86–89.

Croson, Rachel, and Uri Gneezy, 2009, Gender differences in preference, *Journal of Economic Literature* 47, 1–27.

Dana, Jason, Roberto A. Weber, and Jason Xi Kuang, 2007, Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness, *Economic Theory* 33, 67–80.

DellaVigna, Stefan, John A. List, and Ulrike Malmendier, 2012, Testing for altruism and social pressure in charitable giving, *Quarterly Journal of Economics* 127, 1–56.

Deutsch, Morton, and Harold B. Gerhard, 1955, A study of normative and informational influence on individual judgment, *Journal of Abnormal and Social Psychology* 51, 629–636.

Ellingsen, Tore, and Magnus Johannesson, 2008, Pride and prejudice: The human side of incentive theory, *American Economic Review* 98, 990–1008.

Elster, Jon, 1989, Social norms and economic theory, *Journal of Economic Perspectives* 3, 99–117.

Erat, Sanjiv, and Uri Gneezy, 2012, White lies, *Management Science* 58, 723–733.

Fehr, Ernst, and Urs Fischbacher, 2002, Why social preferences matter, *Economic Journal* 112, C1–C33.

——— , 2003, The nature of human altruism, *Nature* 425, 785–791.

Ferraro, Paul J., and Michael K. Price, 2013, Using nonpecuniary strategies to influence behavior: Evidence from a large-scale field experiment, *Review of Economics and Statistics* 95, 64–73.

Fershtman, Chaim, Uri Gneezy, and John A. List, 2012, Equity aversion: Social norms and the desire to be ahead, *American Economic Journal: Microeconomics* 4, 131–144.

Fischbacher, Urs, and Franziska Heusi, 2013, Lies in disguise: An experimental study on cheating, *Journal of the European Economic Association* 11, 525–547.

Fischer, Paul, and Steven Huddart, 2008, Optimal contracting with endogenous social norms, *American Economic Review* 98, 1459–1475.

Fiske, Alan Page, and Philip E. Tetlock, 1997, Taboo tradeoffs: Reactions to transactions that transgress spheres of exchange, *Political Psychology* 17, 255–294.

Gibson, Rajna, Carmen Tanner, and Alexander F. Wagner, 2013, Preferences for truthfulness: Heterogeneity among and within individuals, *American Economic Review* 103, 532–548.

Gino, Francesca, Shahar Ayal, and Dan Ariely, 2009, Contagion and differentiation in unethical behavior, *Psychological Science* 20, 393–398.

Gino, Francesca, Erin L. Krupka, and Roberto A. Weber, 2013, License to cheat: Voluntary regulation and ethical behavior, *Management Science* 59, 2187–2203.

Gneezy, Uri, 2005, Deception: The role of consequences, *American Economic Review* 95, 384–394.

———, Stephan Meier, and Pedro Rey-Biel, 2011, When and why incentives (don't) work to modify behavior, *Journal of Economic Perspectives* 25, 191–210.

Grossman, Zachary, 2012, Self-signaling versus social-signaling in giving, *Working paper.*

———, and Joël van der Weele, 2013, Self-image and strategic ignorance in moral dilemmas, *Working paper.*

Grover, Steven L., 2005, The truth, the whole truth and nothing but the truth: The causes and management of workplace lying, *Academy of Management Executive* 19, 148–157.

Healy, Paul M., and James M. Wahlen, 1999, A review of the earnings management literature and its implications for standard setting, *Accounting Horizons* 13, 365–383.

Huck, Steffen, Dorothea Kübler, and Jürgen Weibull, 2012, Social norms and economic incentives in firms, *Journal of Economic Behavior and Organization* 83, 173–185.

Innes, Robert, and Arnab Mitra, 2013, Is dishonesty contagious?, *Economic Inquiry* 51, 722–734.

Kaptein, Muel, and Mark S. Schwartz, 2007, The effectiveness of business codes: A critical examination of existing studies ad the development of an integrated research model, *Journal of Business Ethics* 77, 111–127.

Kimbrough, Erik O., and Alexander Vostroknutov, 2014, Norms make preferences social, *Working paper, Simon Fraser University and Maastricht University.*

López-Pérez, Raúl, and Eli Spiegelmann, 2013, Why do people tell the truth? Experimental evidence for pure lie aversion, *Experimental Economics* 16, 233–247.

Lundquist, Tobias, Tore Ellingsen, Erik Gribbe, and Magnus Johannesson, 2009, The aversion to lying, *Journal of Economic Behavior and Organization* 70, 81–92.

Luttmer, Erzo F. P., and Monica Singhal, 2014, Tax morale, *Journal of Economic Perspectives* 28, 149–168.

Martin, Steve, 2012, 98% of HBR readers love this article, *Harvard Business Review* pp. 23–25.

Mazar, Nina, On Amir, and Dan Ariely, 2008, The dishonesty of honest people: A theory of self-concept maintenance, *Journal of Marketing Research* 45, 633–644.

Mijovic-Prelec, Danica, and Drazen Prelec, 2010, Self-deception as self-signalling: a model and experimental evidence, *Philosophical Transactions of the Royal Society B* 365, 227–240.

Milgram, Stanley, 1974, *Obedience to Authority* (Harper and Row).

Miller, Nolan, Alexander F. Wagner, and Richard J. Zeckhauser, 2013, Solomonic separation: Risk decisions as productivity indicators, *Journal of Risk and Uncertainty* 46, 265–297.

Ostrom, Elinor, 2000, Collective action and the evolution of social norms, *Journal of Economic Perspectives* 14, 137–158.

Rode, Julian, 2010, Truth and trust in communication - experiments on the effect of a competitive context, *Games and Economic Behavior* 68, 325–338.

Sánchez-Pagés, Santiago, and Marc Vorsatz, 2007, An experimental study of truth-telling in a sender-receiver game, *Games and Economic Behavior* 61, 86–112.

Sherif, Muzafer, 1936, *The Psychology of Social Norms* (Harper and Row: New York).

Sliwka, Dirk, 2007, Trust as a signal of a social norm and the hidden costs of incentive schemes, *American Economic Review* 97, 999–1017.

Tanner, Carmen, Bettina Ryf, and Martin Hanselmann, 2009, Geschützte Werte Skala: Konstruktion und erste Validierung eines Messinstrumentes (Protected Values Measure: Construction and first validation of an instrument to assess protected values), *Diagnostica* 55, 174–183.

Tetlock, Philip E., Orie V. Kristel, S. Beth Elson, Melanie C. Green, and Jennifer S. Lerner, 2000, The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals, *Journal of Personality and Social Psychology* 78, 853–870.

van der Weele, Joël, and Ferdinand von Siemens, 2014, Bracelets of pride and guilt? an experimental test of self-signaling in charitable giving, *CESifo working paper no. 4674*.

Wang, Joseph Tao-Yi, Michael Spezio, and Colin F. Camerer, 2010, Pinocchio's pupil: Using eyetracking and pupil dilation to understand truth-telling and deception in game, *American Economic Review* 100, 984–1007.

Please note: The following supplementary appendices are not meant for publication in print. They can be made available on a Journal website and the authors' websites upon publication.

A.1 Self-signaling model

This Supplementary Appendix provides a model for why one might expect a systematic positive relationship between ρ_i and the agent's type, i.e., $r > 0$. As before, all agents experience morally driven costs of lying, though the amount depends on their type: $C_i = V\theta_i$. Additionally, though, the agent is unsure about (and has imperfect memory of) his type, but can interpret his actions as self-signals of his preferences.

Suppose that there is a continuum of ethical types, distributed continuously with $F(\theta)$ between upper and lower bounds of $\underline{\theta}$ and $\overline{\theta}$, respectively.

Self-signaling is incorporated into the utility function by positing

$$V_{ise}(T) = \begin{cases} -EXCO_{se}\left(1 - r\theta_i\right) + \eta\zeta_1 & \text{if } T = 1 \\ -V\theta_i + \eta\zeta_0 & \text{if } T = 0. \end{cases} \tag{5}$$

Here, ζ_1 is the posterior estimate the agent has about his own type if he tells the truth, ζ_0 is the posterior estimate the agent has about his own type if he lies, and $\eta > 0$ is a parameter which indicates how much the agent cares about his (moral) self-image.

The difference between the utilities of truthtelling and of lying is given by

$$Y^*_{ise} = V\theta_i - EXCO_{se}\left(1 - r\theta_i\right) + \eta\left(\zeta_1 - \zeta_0\right). \tag{6}$$

An individual exhibits truthfulness when $Y^*_{ise} > 0$.

Consider a self-signaling separating equilibrium defined by $\hat{\theta}$ such that for agents with $\theta \geq \hat{\theta}$, $T = 1$ and all other agent types lie. In additional materials available on request, we show that if θ is uniformly distributed over the interval $[\underline{\theta}, \overline{\theta}]$, the cutoff is given by

$$\hat{\theta} = \frac{EXCO_{se} - \frac{\eta\left(\bar{\theta} - \underline{\theta}\right)}{2}}{V + rEXCO_{se}}. \tag{7}$$

We also show that, if $\underline{\theta}$=0, a necessary condition for a unique separating equilibrium of the form postulated to exist is that

$$r > -\frac{V}{EXCO_{se}}. \tag{8}$$

Showing this result formally requires the use of fixed point theorems; the details are available on request. (We also derive similar results for the case when θ is truncated normally distributed.) Intuitively, two agent characteristics support an interpretation of truthfulness as an act of self-signaling. First, when $V > 0$, higher types have higher marginal utility of truthtelling, giving rise to a single-crossing condition. Note, though, that with $V > 0$ high types would be more likely to tell the truth whether or not they engage in self-signaling.

Second, and of primary interest for our paper, the resistance parameter is bounded from below. Sufficiently strong resistance of high types against extrinsic rewards for dishonesty allows self-signaling to work even if being a high type per se does not mean that one values truth as such more. If the direct marginal costs of lying, V, tend to zero,

$r > 0$ is necessary to ensure $\hat{\theta} \geq 0$. The more costly it is to tell the truth, the less the range of possible values for r extends below zero.

A separating equilibrium of the form postulated fails if $EXCO_{se} < 0$; if truthfulness brings benefits, truthtelling is no effective self-signal. Even if $EXCO_{se} > 0$, no equilibrium with $\hat{\theta} \geq 0$ exists if the agent cares too much about self-image, that is, if η is too large. Finally, if r is too small or negative, $\hat{\theta} \leq \overline{\theta}$ may not exist.[1] Of course, in all these cases, people may tell the truth for non-self-signaling reasons.

Overall, $r > 0$ supports a self-signaling equilibrium (though it is not necessary in general).

A.2 Experimental instructions

The instructions of the experiment are attached at the end of this document. This section briefly discusses the construction of the proxies for individual characteristics.

A.2.1 Protected values for truthfulness survey

According to the correspondence (or compatibility) principle established by Ajzen and Fishbein (1980), values and behavior need to be assessed at a similar level of specificity in order to be able to uncover a link between the two. This principle underlies the protected values for truthfulness measure. The questionnaire contains two subscales designed to approach protected values for truthfulness from different angles.

(1) PRV (reactions to violations): Five items assessed the participants' reactions to violations of honesty by a hypothetical CEO who was reporting company information. This scale focuses on the affective dimension of individuals' commitment to honesty.

[*PRV*] Because CEOs' compensation levels depend on the earnings they report to their shareholders, CEOs have an incentive to modify reports to shareholders. What is your opinion on CEOs modifying company information in reports?

Please choose the appropriate category. This is:

Very immoral	1	2	3	4	5	6	7	very moral
Not at all praiseworthy	1	2	3	4	5	6	7	Very praiseworthy
Not at all blameworthy	1	2	3	4	5	6	7	Very blameworthy
Not at all outrageous	1	2	3	4	5	6	7	Very outrageous
Not at all acceptable	1	2	3	4	5	6	7	Very acceptable

[1] If η is large and r or $EXCO_{se}$ are negative, a (pathological) equilibrium where lower types tell the truth can exist.

(2) PNT (no trade): Four additional items assessed the participants' own protected values by examining how much importance they attributed to trade-off reluctance, unwillingness to sacrifice a value, or incommensurability, again referring to the specific context of a hypothetical CEO's decisions regarding the reporting of information.

[*PNT*] CEOs have an opportunity to modify information in the reports they provide to their shareholders. Some view such modification as a violation of truthfulness; others regard it as acceptable protection of personal interests. What do you think about the value of truthfulness in such a situation?

Truthfulness is something

... that one should not sacrifice, no matter what the (material or other) benefits
strongly disagree 1 2 3 4 5 6 7 strongly agree
... for which I think it is right to make a cost-benefit analysis
strongly disagree 1 2 3 4 5 6 7 strongly agree
... that cannot be measured in monetary terms
strongly disagree 1 2 3 4 5 6 7 strongly agree
... about which I can be flexible if the situation demands it
strongly disagree 1 2 3 4 5 6 7 strongly agree

After appropriate recoding of some items, indices of the degrees of protected values for truthfulness were constructed, based on the means across the first five items (for PRV), or the second four items (for PNT). The combined PVT is the mean of all nine items. The original protected values survey was conducted in German. In the paper, for ease of interpretation of the empirical results, we changed the scale to range from 0 to 6. The survey parts were not labeled for participants.

A.2.2 Pro-social concern

We use participants' answers regarding the extent to which they believed that announcing 35 cents had negative consequences for some stakeholders (-2 = hurting some stakeholders to +2 = not hurting some stakeholders) or was manipulative (-2 = manipulative to +2 = not manipulative). Answers to these questions are reordered so that participants who exhibited stronger pro-social concern score high on these scales. We then calculate the mean of the two items. The resulting variable PSC is standardized to have a mean of zero and a standard deviation of unity

A.2.3 Impression management and self-deception

We used the standard Deception Scales (PDS) of Paulhus (1984); see Musch, Brockhaus, and Bröder (2002) for the German version. This is a self-reporting questionnaire designed to measure individuals' tendencies to give socially desirable responses (SDR). See the full instructions for details. It measures two distinct forms of SDR: self-deception and impression management. Accordingly, we coded two variables SELFDECEIT and

IMPRESS. Participants who exhibited more socially acceptable responses scored higher on both scales.

A.2.4 Marginal utility

We asked the following question (drawn from Miller, Wagner, and Zeckhauser (2013)):

Please imagine that you find a CHF 50 bill on the street. It is impossible to identify the owner, and it is, therefore, completely acceptable and morally unobjectionable that you keep the CHF 50. Think about your average peer who earns about the same amount of money as you do, and is approximately equally wealthy. Would you say that, relative to this average peer, you benefit
a lot more
more
equally
less
a lot less
from this additional amount of money?

We assigned a value of 5 to "a lot more" answers, and a value of 1 to "a lot less" answers. This measure captures each participant's self-reported marginal utility of income.

A.3 Additional evidence

Table A.1 documents that demographic characteristics do not interact with situational social norms (see the discussion in Section 4.5).

Table A.1

Situational social norms, economic incentives, and agent-specific costs of lying: Demographics

This Supplementary Appendix table presents coefficients of logit regressions. The dependent variable is TRUTHFUL CHOICE, which is equal to 1 when a participant chose to announce 31 cents and equal to 0 otherwise. Data from the *ECOST*>0 situations are used. Newspaper is a binary indicator equal to 1 if a participant has recently read newspaper articles regarding CEOs (and 0 otherwise). Work is a binary indicator equal to 1 if a participant has a part-time job (and 0 otherwise). Investor is a binary indicator equal to 1 if a participant owns shares, bonds, or mutual funds (and 0 otherwise). All regressions control for all demographic characteristics. The column heading states which of the six characteristics is interacted with *ECOST* and the situational social norm indicators. T-statistics, obtained from robust standard errors clustered at the individual level, appear in parentheses below coefficient estimates. ***, **, and * indicate significance at the 1%, 5%, and 10% levels, respectively.

	(1)	(2)	(3)	(4)	(5)	(6)
Demographic variable used in interactions:	Sex	Age	Economics	Newspaper	Work	Investor
Demographic variable * *APPROVING*	0.00	0.05	0.43	0.19	0.61	0.42
	(0.00)	(0.90)	(0.97)	(0.36)	(1.18)	(0.78)
Demographic variable * *DISAPPROVING*	-0.30	-0.05	-0.05	0.27	-0.22	0.41
	(-0.90)	(-1.43)	(-0.15)	(0.73)	(-0.50)	(1.16)
Demographic variable * *CONTROL*	0.23	-0.01	-0.29	-0.04	-0.80*	0.26
	(0.63)	(-0.33)	(-0.76)	(-0.09)	(-1.94)	(0.67)
Demographic variable * *ECOST*	-0.68**	0.02	0.30	0.49	0.39	0.64*
	(-1.98)	(0.49)	(0.88)	(1.37)	(0.93)	(1.84)
Sex (1: Female, 0: Male)	0.83**	0.36	0.34	0.35	0.37	0.35
	(2.57)	(1.28)	(1.20)	(1.25)	(1.32)	(1.24)
Age (Years)	-0.00	-0.01	-0.01	-0.00	-0.00	-0.01
	(-0.17)	(-0.22)	(-0.20)	(-0.18)	(-0.17)	(-0.19)
Economics (1: Economics, 0: Psychology)	-0.30	-0.28	-0.49	-0.27	-0.27	-0.28
	(-0.99)	(-0.93)	(-1.44)	(-0.91)	(-0.89)	(-0.93)
Newspaper (1: Yes, 0: No)	-0.28	-0.27	-0.28	-0.70**	-0.30	-0.29
	(-0.95)	(-0.92)	(-0.96)	(-2.08)	(-1.03)	(-0.98)
Work (1: Yes, 0: No)	0.29	0.31	0.31	0.30	0.13	0.30
	(1.10)	(1.17)	(1.18)	(1.16)	(0.40)	(1.16)
Investor (1: Yes, 0: No)	-0.03	-0.02	-0.03	-0.04	-0.02	-0.66**
	(-0.12)	(-0.07)	(-0.11)	(-0.14)	(-0.06)	(-2.19)
APPROVING (situational norm, Phase 2)	-0.65**	-1.88	-0.84***	-0.69***	-1.07**	-0.76***
	(-2.08)	(-1.33)	(-2.89)	(-2.72)	(-2.50)	(-2.98)
DISAPPROVING (situational norm, Phase 2)	0.80***	1.89**	0.68***	0.57***	0.83**	0.52**
	(3.36)	(2.14)	(2.91)	(2.85)	(2.09)	(2.45)
CONTROL group (Phase 2)	-0.04	0.31	0.20	0.07	0.67*	-0.02
	(-0.17)	(0.39)	(0.83)	(0.34)	(1.89)	(-0.08)
ECOST (economic cost of no earnings management)	-1.27***	-1.96**	-1.72***	-1.71***	-1.89***	-1.78***
	(-5.93)	(-2.47)	(-7.28)	(-8.32)	(-4.98)	(-8.29)
Constant	0.13	0.39	0.45	0.45	0.45	0.54
	(0.17)	(0.43)	(0.59)	(0.59)	(0.59)	(0.71)
Observations	2,088	2,088	2,088	2,088	2,088	2,088
Pseudo R^2	0.09	0.09	0.09	0.09	0.09	0.09
Pseudo Log Likelihood	-1200	-1201	-1202	-1203	-1198	-1201
Likelihood-ratio test statistic (χ^2, p-value)	240 (<0.01)	238 (<0.01)	236 (<0.01)	234 (<0.01)	244 (<0.01)	238 (<0.01)
Wald test statistic (χ^2, p-value)	111 (<0.01)	102 (<0.01)	106 (<0.01)	106 (<0.01)	104 (<0.01)	108 (<0.01)